Signal Parameter Estimation Using Continuous LASSO and its Performance Analysis

Ashkan Panahi, *Student Member, IEEE,* and Mats Viberg, *Fellow, IEEE,*

***Abstract*—The Least Absolute Shrinkage and Selection Operator (LASSO) has gained attention in a wide class of low dimensional hard estimation problems with promising results. However, such attempts lack strong theoretical background. The objective of this work is to provide a framework for such an analysis. We follow the logical idea of introducing a "best" LASSO estimator covering all possible grid points to avoid the grid choice dependence. We derive conditions for which the so called homotopy method of implementing the proposed Continuous LASSO (CLASS) is consistent in the high SNR scenario assuming infinite estimation dynamic range. We also show that the analysis of CLASS applies to the conventional grid-based LASSO for a sufficiently fine grid. Finally, we give the estimation Mean Squared Error (MSE) for such consistent cases. We present the problem and the numerical results in the context of Direction of Arrival (DOA) estimation.**

***Index Terms*—IEEEtran, journal, LaTeX, paper, template.**

I. Introduction

FINDING a precise sparse linear representation of a signal through the ℓ_1 shrinkage and selection operator has been an active field of study for more than a decade with many applications [1], [2], [3]. The LASSO technique [4], also known as basis pursuit [5], and the global matched filter [6] are among the pioneering such studies. These methods received more attention by different modifications [7], [8], [9], [10] and the invention of effective implementation techniques [11], [12], [13], [14]. There has also been various attempts to formulate and analyze the performance of these methods in some asymptotic cases [15], [16], [17]. The general idea behind such methods is to construct efficient regression vectors to get a better representation which leads to consistency conditions based on less coherent regressors. More recently, there have been different attempts to use LASSO with highly correlated regressors dictated by physical models. These methods have found a wide application in hard low dimensional estimation problems, such as DOA and frequency estimation [18], because of their simple implementation through convex programming techniques [19]. However, these attempts lack a strong theoretical background. In this work, we propose a framework for performing such an analysis.

Framework

We present our work in the context of DOA estimation in the case of far-field sources and narrowband signals, which is a well studied problem [20]. In [18] and [21] the problem is reformulated in a sparse framework and solved using the LASSO technique. However, the theoretical questions about the consistency and the error level of this method is still unanswered. Note also that the analysis can be generalized by replacing the specific terms such as DOAs, waveforms, etc by more general terms such as regressor indexes, regression vector, etc, respectively. For the current purpose, the corresponding terms are equivalently interchangeable.

The main difference of the current study from the previous ones is illustrated by noting the following features:

a) In the current analysis, a distance function between the regressor indexes, a.k.a atom indexes, is defined and we accept a close estimate of the sparsity pattern even if it is not the true one. Consistency means that the distance between the estimated pattern and the true one approaches zero in some asymptotic case, which is the high SNR scenario in this work.

b) Our investigation on the LASSO-based DOA estimator concerns the case in which the Restricted Isometry Property (RIP) [22] does not hold true if the grid is dense enough. However, as we will see, this only results in a fundamental resolution limit on the method. Still, incoherent atoms may be retrieved perfectly even in such a dense grid.

Methodology

Definition: The first obstacle in our analysis is related to the definition of the LASSO estimator. Practically, this method inherits a discretization step which fits the model to that of the ℓ_1 based regression [23]. However, to remove the effect of the grid, one may assume a case of an infinitely small grid size, which transforms the optimization to a functional optimization. In this work we introduce a different framework of estimating a finite set to avoid complications of the functional analysis. However, we show that the proposed continuous LASSO (CLASS) is equivalent to the natural integral optimization extension. We also include some basic necessary properties of the proposed CLASS technique.

Consistency: This work contains the consistency conditions of CLASS in the high SNR case, which corresponds to the so-called noiseless implementation of LASSO. This is generally assumed as a hard method to analyze and to implement. However as proposed in [11], the close connection between the noisy LASSO for small values of Regularization Parameter (RP) and the noiseless LASSO may be exploited to review the properties of the noiseless LASSO. This is generally known as the *homotopy* technique, which we propose as follows:

a) The CLASS cost function is a uniform continuum (homotopy) with respect to the RP between the trivial absolute cost at the infinite RP value and noiseless CLASS at zero.

b) For each value of the RP, the optimal point is unique. Hence, there exists a unique solution component starting from the zero vector and stopping at the CLASS noiseless solution.

c) Under some conditions on DOAs, the solution component is consistent in the sense that for sufficiently small level of noise, there exists a solution point with the correct number of estimates arbitrarily close to the true parameters. It turns out that this point converges to the noiseless solution as the noise vanishes, and the analysis could be done by linearizing in the neighborhood of the noiseless solution. As we see later, there are irregular cases, for which there always exists an additional faulty DOA estimate, but its corresponding waveform value vanishes as the noise tends to zero. In practice, the case might be interpreted as a consistent one, since as noise decreases, eventually the estimated waveform value reaches outside the dynamic range. However, in this work, we stick to the infinite dynamic range assumption neglecting the aforementioned cases.

Performance Analysis: The ultimate goal of this paper is to give the first order performance of CLASS in presence of noise under the consistency conditions. This work will show the best achievable performance of CLASS in terms of bias and Mean Squared Error (MSE), by choosing a proper point in the solution component, given the correct number of sources. The simple approach is to find the irregular or the *singular* point in the path with the given model order ([24]).

Paper Constitution

In the remaining, we introduce CLASS and its optimality conditions in Sections III and III-B respectively. The uniqueness and continuity issues will be discussed in Section IV-A. As we see later in Section IV-D, the best behavior of the CLASS in the high SNR regime is obtained by the near-zero analysis of the regularization parameter, generally known as the *homotopy* rule [11], [24], which we discuss in Section IV-B. To provide a computational condition, we use the CLASS optimality conditions [14] and obtain a consistency rule through the homotopy in Section IV-C. We conclude this paper by giving the asymptotic performance in Section IV-D and showing some results of applying CLASS to the DOA estimation problem in Section V.

II. Problem Statement

We start by a brief review of LASSO-based estimation of DOAs receiving narrowband signals transmitted by far sources.

A. DOA Estimation Problem

Considering a set of n sources at directions $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_n\}$, the received sequence of T sampled signal vectors $\mathbf{X} = [\mathbf{x}(1)\ \mathbf{x}(2) \dots \mathbf{x}(T)]$, at an array of m sensors can be written as

$$\mathbf{X} = \mathbf{A}(\boldsymbol{\theta})\mathbf{S} + \mathbf{N}. \tag{1}$$

where, $\mathbf{S} = [\mathbf{s}(1)\ \mathbf{s}(2) \dots \mathbf{s}(T)]$ and $\mathbf{N} = [\mathbf{n}(1)\ \mathbf{n}(2) \dots \mathbf{n}(T)]$ are the sequences of transmitted data and received noise respectively, and $\mathbf{A}(\boldsymbol{\theta}) = [\mathbf{a}(\theta_1)\ \mathbf{a}(\theta_2) \dots \mathbf{a}(\theta_n)]$ is the collection of the steering vectors corresponding to the DOAs $\boldsymbol{\theta}$. Each steering vector is given by ([6])

$$\mathbf{a}(\theta) = \Big[e^{j\frac{2\pi}{d}r_1\cos(\theta-\rho_1)}\ e^{j\frac{2\pi}{d}r_2\cos(\theta-\rho_2)} \dots e^{j\frac{2\pi}{d}r_m\cos(\theta-\rho_m)}\Big]^T, \tag{2}$$

where (r_i, ρ_i) is the polar coordinate pair of the i^{th} sensor $(i = 1, 2, \dots, m)$ and d is the wavelength at the central frequency. Then, the goal is to estimate $\boldsymbol{\theta}$ given $\mathbf{X}$. Obviously, the problem is defined in a complex-valued space of variables. For simplicity, we focus on the uniform half-wavelength ($r_i = \frac{id}{2}$), linear ($\rho_i = 0$) array. In this case, it is more convenient to write (2) in terms of the electrical angle $\phi = \pi\cos\theta$. We may also use the notation $\mathbf{a}(\phi)$ to show the steering vector as a function of the electrical angle, whenever there is no risk of confusion. We remind that our results are applicable to general estimation problems that can be modeled in the form (1).

B. Conventional LASSO-based Solution

The estimation problem can be solved by first discretizing the parameter space and reformulating the model in the sparse framework [18]. Note that discretization will naturally introduce an additional quantization noise to the estimated parameters. Consider a big set of grid points $\boldsymbol{\theta}_G$ from which we choose the estimated parameters. Let us denote $\mathbf{A}^g = \mathbf{A}(\boldsymbol{\theta}_G)$. Then, the LASSO based estimator is given by

$$\hat{\mathbf{S}}^g(\lambda) = \underset{\mathbf{S}^g}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{X} - \mathbf{A}^g\mathbf{S}^g\|_F^2 + \lambda\|\mathbf{S}^g\|_{1,2}, \tag{3}$$

where, for every $\mathbf{S}$ with S_{ij} as the element in the i^{th} row and j^{th} column,

$$\|\mathbf{S}\|_{1,2} = \sum_i \gamma_{i,\mathbf{S}}, \tag{4}$$

with

$$\gamma_{i,\mathbf{S}} = \sqrt{\sum_j |S_{ij}|^2}, \tag{5}$$

and $\boldsymbol{\gamma}_{\mathbf{S}} = [\gamma_{1,\mathbf{S}}\ \gamma_{i,\mathbf{S}} \dots \gamma_{n,\mathbf{S}}]$. We also refer to the diagonal matrix of the elements in $\boldsymbol{\gamma}_{\mathbf{S}}$ as $\boldsymbol{\Gamma}_{\mathbf{S}}$. We further introduce $\hat{\boldsymbol{\theta}}^g(\lambda) \subset \boldsymbol{\theta}_G$ as the set of all DOAs corresponding to the nonzero rows in $\hat{\mathbf{S}}(\lambda)$. We will refer to (3) as $\mathscr{P}_g$ optimization. As explained in [4], the estimation can be performed in the noiseless case by solving

$$\hat{\mathbf{S}}^g_{\text{nl}} = \underset{\mathbf{S}^g}{\operatorname{argmin}} \|\mathbf{S}^g\|_{1,2} \text{ subject to } \mathbf{A}^g\mathbf{S}^g = \mathbf{X}, \tag{6}$$

with the active bases corresponding to directions in $\hat{\boldsymbol{\theta}}^g_{\text{nl}}$. We will refer to this optimization as $\mathscr{P}_{g,\text{nl}}$.

III. Continuous LASSO

A natural method of extending the above technique to estimate continuous DOA parameters is to let the grid size increase so that $\mathbf{A}^g$ covers all possible directions. In this case, one may assume that the summations tend to proper integrals. However, due to technical difficulties we avoid such a method, and instead introduce the following optimization which leads to an extension of $\mathscr{P}_g$. In the end of this section, we show the equivalence of this method to the former integral extension.

A. CLASS construction

The CLASS construction is based on the observation that the LASSO solution is always sparse. Denoting

$$P(\mathbf{X},\boldsymbol{\theta},\lambda)=\min_{\mathbf{S}}\Psi(\mathbf{X},\boldsymbol{\theta},\mathbf{S},\lambda), \tag{7}$$

where

$$\Psi(\mathbf{X},\boldsymbol{\theta},\mathbf{S},\lambda)=\frac{1}{2}\left\|\mathbf{X}-\mathbf{A}(\boldsymbol{\theta})\mathbf{S}\right\|_F^2+\lambda\|\mathbf{S}\|_{1,2}, \tag{8}$$

it is simple to check that $P(\mathbf{X},\boldsymbol{\theta},\lambda)$ exists for each $\lambda\geq 0$ and it is a continuous function of $(\boldsymbol{\theta},\lambda)\in[0\ 2\pi]^n\times\mathbb{R}$ for each $n\in\mathbb{N}$. Moreover, (7) is an extension of (3). In fact, (3) can easily be obtained from (7) by substituting $\boldsymbol{\theta}=\boldsymbol{\theta}_G$. We call a DOA set $\boldsymbol{\theta}$ *reducible* at λ if there exists a global minimum point $\hat{\mathbf{S}}$ of (7), at least one row of which is completely zero. In this case, $\boldsymbol{\theta}$ can be *reduced* to a $\boldsymbol{\theta}'\subset\boldsymbol{\theta}$ by removing the DOAs corresponding to the zero rows. Note that if $\boldsymbol{\theta}$ can be reduced to $\boldsymbol{\theta}'$ at λ, then $P(\mathbf{X},\boldsymbol{\theta},\lambda)=P(\mathbf{X},\boldsymbol{\theta}',\lambda)$. The final observation is that if the dimension n of $\boldsymbol{\theta}$ is greater than $2m$ it is reducible, due to the convex conic nature of the optimization. Next, we define

$$\begin{aligned}\hat{\boldsymbol{\theta}}(\mathbf{X},\lambda)&=\underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}P(\mathbf{X},\boldsymbol{\theta},\lambda)\\&\text{subject to }\boldsymbol{\theta}\text{ is irreducible at }\lambda,\end{aligned} \tag{9}$$

where $\Theta=\bigcup_{k=1}^{\infty}[0\ \pi]^k$ is the algebra of all finite subsets of the interval $[0\ \pi]$. Note that due to the constraint, the solution dimension is bounded by $2m$ and $\hat{\boldsymbol{\theta}}$ always exists, since the optimization can be confined to a finite union of compact sets. We refer to (9) as $\mathscr{P}_t$ optimization. We can also define the noiseless continuous optimization $\mathscr{P}_{t,\mathrm{nl}}$ by

$$\{\hat{\boldsymbol{\theta}}_{\mathrm{nl}},\hat{\mathbf{S}}_{\mathrm{nl}}\}=\underset{\boldsymbol{\theta}\in\Theta,\ \mathbf{S}}{\operatorname{argmin}}\|\mathbf{S}\|_1\text{ subject to }\mathbf{A}(\boldsymbol{\theta})\mathbf{S}=\mathbf{X}. \tag{10}$$

The goal of this paper is to discuss the performance of this continuous version of LASSO in the asymptotic low noise scenario of the DOA data model. This will be mainly discussed in Section IV. The consistency of the $\mathscr{P}_g$ optimization will also be presented. We finish this section by stating some basic observations and introducing some new definitions in the noiseless scenario which will be necessary for the analysis.

B. Preliminary Observations

This section contains two main results. First, we rationalize the idea of CLASS by comparing it with another natural extension. Next, we give a set of optimality conditions for $\mathscr{P}_t$. We start by discussing a straightforward extension of (3) by letting the grid size increase so that the sums tend to integrals. On the limit, the vector $\mathbf{S}^g$ tends to a vector measure $\mathbf{s}:\mathscr{B}_{[0\ \pi]}\rightarrow\mathbb{C}^T$, where $\mathscr{B}_{[0\ \pi]}$ is the restriction of the real Borel σ-algebra to $[0\ \pi]$. Then, we can show the following.

Theorem 1. For a certain realization of $\mathbf{X}$, assume that the $\mathscr{P}_t$ solution is given by $\hat{\boldsymbol{\theta}}(\mathbf{X},\lambda)$. Then,

$$P(\mathbf{X},\hat{\boldsymbol{\theta}}(\mathbf{X},\lambda),\lambda)=\min_{\mathbf{s}}\frac{1}{2}\left\|\mathbf{X}-\int_{[0\ \pi]}\mathbf{a}(\theta)\mathrm{d}\mathbf{s}^H\right\|_F^2+\lambda\|\mathbf{s}\|, \tag{11}$$

where the minimum is taken over all T-dimensional complex vector measures on $\mathscr{B}_{[0\ \pi]}$ and $\|\mathbf{s}\|$ denotes the total variation of $\mathbf{s}$ over $[0\ \pi]$.

Proof: The proof is given in [25]. ∎

Now, we give a general characterization of the global minimum of $\mathscr{P}_t$. Indeed, this plays a central role in the later analysis which includes some steps as follows.

Theorem 2. Suppose $\mathbf{S}$ is an $n\times T$ matrix with $\mathbf{S}_i$ as the i^{th} row and $\boldsymbol{\theta}=[\theta_1\ \theta_2\ \ldots\theta_n]$. By introducing $\hat{\mathbf{N}}=\mathbf{X}-\mathbf{A}(\boldsymbol{\theta})\mathbf{S}$, the matrix $\mathbf{S}$ is a global minimum in (7) if and only if first,

$$\mathbf{a}^H(\theta_i)\hat{\mathbf{N}}=\lambda\gamma_{i,\mathbf{S}}^{-1}\mathbf{S}_i, \tag{12}$$

for every θ_i that corresponds to an active row in $\mathbf{S}$, and second,

$$\|\mathbf{a}^H(\theta_i)\hat{\mathbf{N}}\|_2\leq\lambda, \tag{13}$$

for $i=1,2,\ldots,n$.

Proof: See [14] for a proof. ∎

For the sake of simplicity, we denote the row in $\mathbf{S}$ corresponding to the DOA θ in $\hat{\boldsymbol{\theta}}$ by $\mathbf{S}_\theta$. From Theorem 2 two basic results can be inferred. We explain them in the following two corollaries.

Corollary 1. CLASS optimality condition. The matrix $\mathbf{S}$ and the DOA vector $\boldsymbol{\theta}$ in Theorem 2 form a global minimum for (9) if and only if $\mathbf{S}$ is irreducible and the following conditions are satisfied:

$$\begin{aligned}&\mathbf{A}^H(\boldsymbol{\theta})\hat{\mathbf{N}}=\lambda\boldsymbol{\Gamma}_{\mathbf{S}}^{-1}\mathbf{S}\\&\max_{0\leq\theta\leq\pi}\|\mathbf{a}^H(\theta)\hat{\mathbf{N}}\|_2\leq\lambda,\end{aligned} \tag{14}$$

Proof: First, assume that $\mathbf{S}$ is an irreducible global minimum for the $\mathscr{P}_t$. Then, for each DOA θ_0 with $0\leq\theta_0\leq\pi$, the vector $\mathbf{S}_1=[\mathbf{S}^H\ \mathbf{0}]^H$ is a global minimum of (7) with $\boldsymbol{\theta}_1=[\boldsymbol{\theta}\ \theta_0]$. Using Theorem 2 we conclude (14). Now, suppose that $\mathbf{S}$ and $\boldsymbol{\theta}$ satisfy (14). For each different DOA set $\boldsymbol{\theta}_1$, introducing $\boldsymbol{\theta}_2=\boldsymbol{\theta}_1\cup\boldsymbol{\theta}$ we observe that $\mathbf{S}_2=[\mathbf{S}^H\ \mathbf{0}]^H$ satisfies the conditions in Theorem 2. Thus, $\boldsymbol{\theta}_2$ is reducible to $\boldsymbol{\theta}$ and $P(\mathbf{X},\boldsymbol{\theta}_2,\lambda)=P(\mathbf{X},\boldsymbol{\theta},\lambda)$. On the other hand because $\boldsymbol{\theta}_1\subset\boldsymbol{\theta}_2$ we have $P(\mathbf{X},\boldsymbol{\theta}_2,\lambda)\leq P(\mathbf{X},\boldsymbol{\theta}_1,\lambda)$. Putting the results together, we complete the proof. ∎

Corollary 2. Second order sufficiency. The $\mathscr{P}_t$ solution and the solution to (7) are only functions of the observed sample correlation matrix $\mathbf{R}_x=\frac{1}{T}\mathbf{X}\mathbf{X}^H$.

Proof: We only need to show that the solution to $\mathbf{S}$ in (12) is a function of $\mathbf{R}_x$. This can easily be shown by substituting the definition of $\hat{\mathbf{N}}$ in Theorem 2 into (12) to obtain

$$\mathbf{S}=\left(\mathbf{A}^H\mathbf{A}+\lambda\boldsymbol{\Gamma}_{\mathbf{S}}\right)^{-1}\mathbf{X}, \tag{15}$$

which solves $\mathbf{S}$ with respect to $\boldsymbol{\Gamma}_{\mathbf{S}}$. According to the fact that $\boldsymbol{\Gamma}_{\mathbf{S}}$ consists of the diagonal elements of $\mathbf{S}\mathbf{S}^H$, and substituting (15), it is obvious that $\boldsymbol{\Gamma}_{\mathbf{S}}$ and, consequently, $\mathbf{S}$ are functions of $\mathbf{R}_x$ only. ∎

As a result of Corollary 2, we may assume that $\mathbf{X}$ is of full column rank. Otherwise, there exists a full rank matrix $\mathbf{X}'$ with fewer columns and the same sample correlation matrix. However, to maintain the statistical properties of $\mathbf{X}$, we avoid such a restriction in this stage.

A similar set of results can be obtained for the noiseless optimization $\mathscr{P}_{t,\text{nl}}$, which is the modified version of the results, such as the null space property originally given in [26], on which the RIP theory of perfect recovery is based. For a complete analysis see [25].

IV. Performance Analysis

A. General properties of the solution path

We start this section by a nice result showing that the CLASS optimization over the ULA manifold is unique. Although unproved, we also claim that this theorem is valid for all unambiguous array manifolds based on pure empirical observations.

Theorem 3. The CLASS solution uniqueness. The solutions to $\mathscr{P}_{t,\text{nl}}$ and $\mathscr{P}_t$, for every $\lambda > 0$, are unique and, at most, of dimension $n = m - 1$.

Proof: The proof is given in Appendix A. ∎

As a remark, note that the global sparsity upper bound of $n = m - 1$ is not obvious in the complex case. The reader may verify that the resulting sparse vector from complex LASSO by a random regression matrix can be nonzero in possibly $2m - 1$ elements. This tighter bound reflects the unique characteristics of the array manifold. The immediate consequence of Theorem 3 is that the unique solutions of $\mathscr{P}_t$ and $\mathscr{P}_{t,\text{nl}}$ are continuous with respect to all their variables ($\mathbf{X}$ and λ). For the current purpose, we focus on the continuum at $\lambda = 0$. However, due to multidimensionality, the closeness definition should be clarified on Θ. A general expectation is that two close solution pairs $(\boldsymbol{\theta}_i, \mathbf{S}_i)$ with $i = 1, 2$ get close in a sense that their elements corresponding to "far" DOAs gradually vanish. The $\mathscr{P}_t$ continuum may also extend to $\lambda = 0$, in which case it approaches the $\mathscr{P}_{t,\text{nl}}$ solution. This is generally known as the *homotopy* principle in the literature [24]. The following theorem summarizes and clarifies these ideas.

Theorem 4. Solutions to $\mathscr{P}_t$ and $\mathscr{P}_{t,\text{nl}}$ are continuous, i.e. for every $\lambda > 0$, $\mathbf{X}$, and $\epsilon > 0$, there exists $\delta > 0$ so that if $\max\{\|\Delta\mathbf{X}\|, \|\Delta\lambda\|\} < \delta$, then for all $\theta \in \hat{\boldsymbol{\theta}}(\mathbf{X}+\Delta\mathbf{X}, \lambda+\Delta\lambda)$ either there exists a $\theta' \in \hat{\boldsymbol{\theta}}(\mathbf{X}, \lambda)$ so that $\max\{|\theta - \theta'|, \|\mathbf{S}_\theta - \mathbf{S}_{\theta'}\|\} < \epsilon$ or $\|\mathbf{S}_{\theta'}\| < \epsilon$.

Furthermore, as λ tends to zero, the solution to $\mathscr{P}_t$ tends to a solution of $\mathscr{P}_{t,\text{nl}}$ in the above sense.

Proof: The proof follows from a standard procedure. However, due to the different continuum definition we give a sketch of it in Appendix B . ∎

As we have already stated, the continuity of the LASSO path at $\lambda = 0$ is widely referred to as the homotopy property.

B. The near-zero behavior of the LASSO path

The homotopy rule also enables us to perform the analysis for very small values of λ and small deviations $\Delta\mathbf{X}$ from the noiseless measurement $\mathbf{X}$, in which case we may utilize the first order expansion tools. As previously explained, we use the notation $\mathbf{a}(\phi)$ to show the steering vector as a function of the electrical angle in the ULA case. We confine the analysis for the cases in which, fixing $\mathbf{X}$, the dimension of the solution is equal to the dimension of the noiseless solution in a sufficiently small neighborhood of $\lambda = 0$. We may call such case a *pure* case. We also denote $\mathbf{d}(\theta) = \frac{\mathrm{d}\mathbf{a}(\theta)}{\mathrm{d}\theta}$ and $\mathbf{D}(\boldsymbol{\theta}) = [\mathbf{d}(\theta_1)\ \mathbf{d}(\theta_2) \dots \mathbf{d}(\theta_n)]$. As we have already shown by continuity, for sufficiently small values of λ and $\Delta\mathbf{X}$, the solution is close to the noiseless one. Let us say that the solution pair at such a point is given by $(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}, \mathbf{S} + \Delta\mathbf{S})$, where $(\boldsymbol{\theta}, \mathbf{S})$ is the solution of the $\mathscr{P}_{t,\text{nl}}$. We recall that we neglect the possibility of the existence of additional vanishing DOAs. Expanding the conditions in Theorem 2, up to the second order, we get

$$\left\|\mathbf{a}^H(\theta)\left(\Delta\mathbf{X} - \mathbf{A}(\boldsymbol{\theta})\Delta\mathbf{S} - \mathbf{D}(\boldsymbol{\theta})\Delta Q'\mathbf{S}\right)\right\|_2 \leq \lambda \tag{16}$$

$$\mathbf{a}^H(\theta_i)\left(\Delta\mathbf{X} - \mathbf{A}(\boldsymbol{\theta})\Delta\mathbf{S} - \mathbf{D}(\boldsymbol{\theta})\Delta Q\mathbf{U}\right) = \lambda\mathbf{U}_i, \tag{17}$$

where we introduced $\mathbf{U} = \boldsymbol{\Gamma}_{\mathbf{S}}^{-1}\mathbf{S}$, with $\mathbf{U}_i$ as the i^{th} row, and $\Delta Q = \Delta Q'\boldsymbol{\Gamma}$ with $\Delta Q'$ as the diagonal matrix whose diagonal elements are identical to $\Delta\boldsymbol{\theta}$. Note that $\|\mathbf{U}_i\|_2 = 1$.

We observe from (16) that θ_i is a maximum point for the function $f(\theta_0) = \|\mathbf{a}^H(\theta_0)\left(\Delta\mathbf{X} - \mathbf{A}(\boldsymbol{\theta})\Delta\mathbf{S} - \mathbf{D}(\boldsymbol{\theta})\Delta Q\mathbf{U}\right)\|_2$. Thus, its first derivative is equal to zero and we get

$$\mathrm{Re}\left(\mathbf{d}^H(\theta_i)\left(\Delta\mathbf{X} - \mathbf{A}(\boldsymbol{\theta})\Delta\mathbf{S} - \mathbf{D}(\boldsymbol{\theta})\Delta Q\mathbf{U}\right)\mathbf{U}_i^H\right) = 0. \tag{18}$$

We can solve (18) and (17) to obtain the first order expansions. Let us define $\mathbf{P}_{\mathbf{A}}^{\perp} = \mathbf{I} - \mathbf{A}(\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H$ as the orthogonal projection matrix of the operator $\mathbf{A}$, the matrix $\boldsymbol{\Xi} = \mathbf{U}\mathbf{U}^H$ with $\boldsymbol{\xi}_i$ as the i^{th} column, and $\mathbf{R} = \mathrm{Re}\left[\mathbf{D}_p^H\mathbf{D}_p \odot \boldsymbol{\Xi}^T\right]$, where $\mathbf{D}_p = \mathbf{P}_{\mathbf{A}}^{\perp}\mathbf{D}$ and $\odot$ denotes the element wise multiplication. Then, we have $\Delta\boldsymbol{\theta} = \boldsymbol{\delta} + \lambda\boldsymbol{\beta}$ where

$$\boldsymbol{\beta} = \boldsymbol{\Gamma}_{\mathbf{S}}^{-1}\mathbf{R}^{-1}\mathrm{Re}\begin{bmatrix} \mathbf{d}_1^H\mathbf{A}(\mathbf{A}^H\mathbf{A})^{-1}\boldsymbol{\xi}_1 \\ \vdots \\ \mathbf{d}_n^H\mathbf{A}(\mathbf{A}^H\mathbf{A})^{-1}\boldsymbol{\xi}_n \end{bmatrix} \tag{19}$$

$$\boldsymbol{\delta} = \boldsymbol{\Gamma}_{\mathbf{S}}^{-1}\mathbf{R}^{-1}\mathrm{Re}\begin{bmatrix} \mathbf{d}_1^H\mathbf{P}_{\mathbf{A}}^{\perp}\Delta\mathbf{X}\mathbf{U}_1^H \\ \vdots \\ \mathbf{d}_n^H\mathbf{P}_{\mathbf{A}}^{\perp}\Delta\mathbf{X}\mathbf{U}_n^H \end{bmatrix}, \tag{20}$$

where we dropped the arguments of $\mathbf{A}(\boldsymbol{\theta})$ and $\mathbf{D}(\boldsymbol{\theta})$ for simplicity. Further, we get $\Delta\mathbf{S} = (\mathbf{A}^H\mathbf{A})^{-1}\left[\mathbf{A}^H(\Delta\mathbf{X} - \mathbf{D}\Delta Q\mathbf{U}) - \lambda\mathbf{U}\right]$. We recognize (20) as the first order expansion of the standard NLLS estimation error [27], whereas (19) represents the contribution due to the ℓ_1 regularization.

C. Consistency Analysis

In this section, we discuss the correctness of the solution in the asymptotic noiseless case. It turns out that the LASSO based estimation is not generally consistent. Our work is based on the definitions given in [28], where an asymptotic consistency resolution lower bound of $\Delta\phi = \frac{\pi}{m}$ is computed for the case of ULAs by trying the null space property on a set of candidate null vectors and the discretized LASSO based estimates. Likewise, we focus on the case $n = 2$ and look for a condition on the separation of the two true DOAs, under which the LASSO based estimator gives correct result irrespective of the true source vector $\mathbf{S}$. As a comparison, note that there is no such consistent case for the classical beamforming technique [20], which cannot resolve sources with too different power levels. Also, note that to investigate the consistency one should consider the case of infinitesimal noise rather than the noiseless one. As usual, there is a clear relation between the two cases given by the homotopy rule. Thus, we start by the noiseless case ($\Delta\mathbf{X} = \mathbf{0}$) and look for the so called *perfect recovery* bounds. Finally, note that in the case of there exists a simple symmetry in linear arrays. It is simple to see that fixing the true $\mathbf{S}$, a rigid shift in true DOA electrical angles results in a rigid shift in the CLASS solution. Thus, we may take two DOAs $\theta = \frac{\pi}{2}, \frac{\pi}{2} + \Delta\theta$ corresponding to $\phi = 0, \Delta\phi$ respectively without loss of generality if the array is uiform and linear.

We use the condition (16) with the results in Section IV-B to obtain an upper bound. If the noiseless solution is identical to the true solution, (16) should be correct with the true parameters and the ones computed in (19). One may show that (16) is also a sufficient condition for the pure cases. Note that there may exist consistent impure cases, which obviously do not satisfy (16). Thus, what we obtain is an upper bound for a more general framework, in which there is a finite dynamic range. Since $\Delta\mathbf{X} = \mathbf{0}$, we have $\boldsymbol{\delta} = \mathbf{0}$. Then, $\Delta\boldsymbol{\theta} = \lambda\boldsymbol{\beta}$ and $\Delta\mathbf{S} = -\lambda(\mathbf{A}^H\mathbf{A})^{-1}\left[\mathbf{A}^H\mathbf{D}\Delta Q_0 + \mathbf{I}\right]\mathbf{U}$, where ΔQ_0 is the diagonal matrix of the elements of $\boldsymbol{\Gamma}_{\mathbf{S}}\boldsymbol{\beta}$. We can further write (16) as follows

$$\forall\theta,\ \|\mathbf{a}^H(\theta)\left(\mathbf{P}_{\mathbf{A}}^{\perp}\mathbf{D}\Delta Q_0 - \mathbf{A}(\mathbf{A}^H\mathbf{A})^{-1}\right)\boldsymbol{\Xi}^{\frac{1}{2}}\|_2 \leq 1 + O(\lambda), \tag{21}$$

Letting λ tend to zero, we observe that (21) is also a zero order necessary condition removing the term $O(\lambda)$. This means that

Result 1. For any noiseles solution of CLASS, we have

$$\forall\theta,\ \|\mathbf{a}^H(\theta)\left(\mathbf{P}_{\mathbf{A}}^{\perp}\mathbf{D}\Delta Q_0 - \mathbf{A}(\mathbf{A}^H\mathbf{A})^{-1}\right)\boldsymbol{\Xi}^{\frac{1}{2}}\|_2 \leq 1. \tag{22}$$

For a certain m, at a DOA set $\boldsymbol{\theta}$, if the condition (22) is true for every θ and a certain choice of $\boldsymbol{\Xi}$, this DOA set is said to be consistent over $\boldsymbol{\Xi}$. Note that ΔQ is also a function of $\boldsymbol{\Xi}$. We can also simplify the analysis by focusing on the asymptotic case of very large number of sensors m. A more general terminology is given as follows.

Definition 1. For a fixed m, a DOA set $\boldsymbol{\theta}$ is said to be *absolutely consistent* if (22) is true for every choice of $\boldsymbol{\Xi}$. A DOA set is called *almost consistent* if (22) is true for almost all matrices $\boldsymbol{\Xi}$. Furthermore, a sequence $\boldsymbol{\theta}_m$ is called *asymptotically consistent* if for every $\boldsymbol{\Xi}$ there exists a large number M such that for $m > M$, the set $\boldsymbol{\theta}_m$ is consistent over $\boldsymbol{\Xi}$. Finally, a sequence is *almost asymptotically consistent* if the later condition is true for almost all matrices $\boldsymbol{\Xi}$.

Asymptotic Consistency for ULA

Using these definitions, we provide some asymptotic consistency conditions when the array is uniform and linear. We first need to introduce the functions $F_a(\delta) = \frac{e^{j\delta}-1}{j\delta}$, $G_a(\delta) = \frac{dF_a}{d\delta}$, and $H_a(\delta) = -\frac{dG_a}{d\delta}$. We also define the 2×2 matrices $\mathbf{F}_a(\delta) = \begin{bmatrix} 1 & F_a(\delta) \\ F_a(-\delta) & 1 \end{bmatrix}$, $\mathbf{G}_a(\delta) = \begin{bmatrix} \frac{1}{2}j & G_a(\delta) \\ G_a(-\delta) & \frac{1}{2}j \end{bmatrix}$, $\mathbf{H}_a(\delta) = \begin{bmatrix} \frac{1}{3} & H_a(\delta) \\ H_a(-\delta) & \frac{1}{3} \end{bmatrix}$, and the 1×2 vectors $\mathbf{g}(\delta',\delta) = [G(\delta)\ G(\delta+\delta')]$ and $\mathbf{f}(\delta',\delta) = [F(\delta)\ F(\delta+\delta')]$.

Neglecting the details, when $m \to \infty$, (22) leads to

Result 2. assuming a sequence $\boldsymbol{\phi}_m = [0\ \Delta\phi_m]$,

- $\boldsymbol{\phi}_m$ is almost asymptotically consistent if $m\Delta\phi_m \to \infty$.
- It is not almost asymptotically consistent if $\Delta\phi_m = o(\frac{1}{m})$.
- If $\Delta\phi_m \sim \frac{\delta}{m}$ with a fixed δ the sequence is almost asymptotically consistent if for every $\delta' \in \mathbb{R}$

$$\|(\mathbf{g}_a(\delta',\delta) - \mathbf{f}_a(\delta',\delta)\mathbf{F}_a^{-1}\mathbf{G}_a)\Delta Q_a - \mathbf{f}_a(\delta',\delta)\mathbf{F}_a^{-1}\|_2 \leq 1 \tag{23}$$

Result 2 shows that in the above asymptotic case, the inconsistency is reflected by the resolution limit of the form $\frac{\delta}{m}$, where δ is the smallest number satisfying (23). In Section V we use (23) and (22) to find out a threshold rate in the asymptotic case and the resolution of small arrays respectively.

Discretizing Consistency

Finally, in this section we review the behavior of the discretized LASSO based estimator $\mathscr{P}_g$ for a very fine grid of DOAs. We define $\Delta(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \max_{\theta_1\in\boldsymbol{\theta}_1}\min_{\theta_2\in\boldsymbol{\theta}_2}|\theta_1-\theta_2|$ as the asymmetric distance between two DOA sets $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. For each grid $\boldsymbol{\theta}$, we also define $\zeta(\boldsymbol{\theta}) = \max_{\theta\in[0\ \pi]}\min_{\theta'\in\boldsymbol{\theta}}|\theta-\theta'|$ as the measure of fineness.

Theorem 5. Discretizing consistency. Suppose that the $\mathscr{P}_t$ minimum point is given by $\boldsymbol{\theta}_0$. Then, for each $\epsilon > 0$ there exists a $\delta > 0$ so that if $\Delta(\boldsymbol{\theta}_0,\boldsymbol{\theta}) \leq \delta$ then $\boldsymbol{\theta}$ can be reduced to a DOA set $\boldsymbol{\theta}'$ so that $\Delta(\boldsymbol{\theta}',\boldsymbol{\theta}) \leq \epsilon$.

Proof: The proof is straightforward noting the uniqueness theorem. Assume that the theorem is not true. Then, there must exist an ϵ and a sequence of DOAs $\boldsymbol{\theta}_n$ reducible to $\boldsymbol{\theta}'_n$ such that $\Delta(\boldsymbol{\theta}_0,\boldsymbol{\theta}_n) \leq \frac{1}{n}$ but $\Delta(\boldsymbol{\theta}'_n,\boldsymbol{\theta}) > \epsilon$. It is then simple to see that this leads to a $\mathscr{P}_t$ minimum point different than $\boldsymbol{\theta}_0$ which is in contradiction with uniqueness. ■

Theorem 5 shows that if the grid $\boldsymbol{\theta}_G$ in $\mathscr{P}_g$ satisfies $\zeta(\boldsymbol{\theta}_G) < \delta$ with a proper δ then the active basis of the $\mathscr{P}_g$ solution are ϵ close to the true solution of $\mathscr{P}_t$. Thus, the analysis of CLASS applies to the grid-based LASSO for a sufficiently fine grid.

D. Asymptotic Estimator Variance

In this section, we complete the LASSO based DOA estimation performance analysis by giving the asymptotic performance of the CLASS estimator in the low noise case. In this case, the noiseless solution will not give the right number of DOAs and the $\mathscr{P}_t$ optimization with $\lambda > 0$ should be utilized to compensate. Assume that the true DOA set $\boldsymbol{\theta}$ is pure and consistent. The noiseless solution parameters are equal to the true ones. Thus, we may use the true parameters using (19). In this case, $\boldsymbol{\gamma}_{i,\mathbf{S}}$ is the sample Root Mean Square (RMS) of the i^{th} source and $\boldsymbol{\Xi}$ is the sample correlation coefficient matrix of the sources. The parameter λ can be computed as the smallest value for which the computed perturbations satisfy (16). Recall that the consistency guarantees that for sufficiently small $\Delta\mathbf{X} = \mathbf{N}$ such a λ value exist. We assume a white, Gaussian, centered, circularly symmetric noise vector process $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ and $\mathbf{N} = [\mathbf{n}_1\ \mathbf{n}_2 \dots \mathbf{n}_T]$. According to the results in Section IV-B, we have $\Delta\boldsymbol{\theta} = \boldsymbol{\delta} + \lambda\boldsymbol{\beta}$ where $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ are given in (20) and (19) respectively. Recall that $\boldsymbol{\delta}$ is actually the asymptotic Maximum Likelihood (ML) error in high SNR scenario [27].

Result 3. The asymptotic statistical DOA bias is then given by

$$\mathcal{E}(\Delta\boldsymbol{\theta}) = \mathcal{E}(\lambda)\boldsymbol{\beta}, \tag{24}$$

where we used the fact that $\mathcal{E}(\boldsymbol{\delta}) = 0$. The asymptotic error covariance is

$$\begin{aligned}&\text{Cov}(\Delta\boldsymbol{\theta}) = \\ &\tfrac{1}{2}\boldsymbol{\Gamma}^{-1}\mathbf{R}^{-1}\text{Re}[\mathbf{D}_p^H\mathbf{C}\mathbf{D}_p \odot \boldsymbol{\Xi}]\mathbf{R}^{-1}\boldsymbol{\Gamma}^{-1} + \text{Var}(\lambda)\boldsymbol{\beta}\boldsymbol{\beta}^T,\end{aligned} \tag{25}$$

where more details are presented in Appendix C.

As seen, the error statistics are affected by the choice of λ. We remind that the strategy we follow here is to choose λ such that CLASS gives the correct number of sources, which is possible only for pure cases in high SNR. Due to (25), in terms of minimum error, the best choice is the smallest such λ. Thus, using (16), we get

$$\begin{aligned}&\lambda = \min \lambda' \\ &\text{such that} \\ &\forall\theta,\ \lambda' \geq \|\mathbf{a}^H(\theta)\{\mathbf{P}_{\mathbf{A}}^{\perp}(\mathbf{N} - \mathbf{D}\Delta Q\mathbf{U}) + \lambda'\mathbf{A}(\mathbf{A}^H\mathbf{A})^{-1}\mathbf{U}\}\|.\end{aligned} \tag{26}$$

Let us denote the set of all non-negative values of λ satisfying the inequality in (26) for a certain θ by S_θ. Then $\lambda = \min \bigcap_\theta S_\theta$. Note that for each $\theta \in \boldsymbol{\theta}_0$, $S_\theta = [0\ \infty)$ while for other values $S_\theta = [\Lambda(\theta)\ \infty)$, where $\Lambda(\theta)$ can be found by solving the third line of (26) with equality. Now, obviously, $\lambda = \max_\theta \Lambda(\theta)$ (see a similar argument in [14] for more illustration). For large enough m, the relation (26) can be simplified. We assume that $\Lambda(\theta)$ for different values of θ is almost at the same level so that $\theta_1 = \arg\max \Lambda(\theta)$ is well-distributed such that $\Pr(\min_{\theta\in\boldsymbol{\theta}_0} |\theta_1 - \theta| = O(\frac{1}{m})) = 0$, where $\boldsymbol{\theta}_0$ is the true DOA electrical angle set. This means that $\mathbf{a}^H(\theta_1)\mathbf{A} = o(m)$, $\mathbf{a}^H(\theta_1)\mathbf{D} = o(m^2)$ and, $\mathbf{a}^H(\theta_1)\mathbf{P}_{\mathbf{A}}^{\perp}$ is identical to $\mathbf{a}^H(\theta_1)$ almost surely. Following this line of reasoning, after some manipulations and keeping the dominant terms, we get

$$\lambda \approx \max_\theta \|\mathbf{a}^H(\theta)\mathbf{N}\|_2. \tag{27}$$

Even with such a simplification, finding the two first statistical moments of λ is a hard task. However, for the case of uncorrelated noise, $\mathbf{C} = \sigma^2\mathbf{I}$, and large enough m in a ULA, we can proceed further by focusing on the extreme value, λ_s^2 of the sampled set $\left\{z_k = \|\mathbf{a}^H(\frac{2k\pi}{m})\mathbf{N}\|_2^2\right\}_{k=1}^m$. In this case,

$$\frac{\lambda_s^2}{\sigma^2 m} = \ln m + (T-1)\ln\ln m + \gamma - \ln(T-1)!, \tag{28}$$

where γ converges weakly to a Gumble [29] random variable. It should be also reminded that this result is valid when $T = o(\frac{\ln m}{\ln\ln m})$. See Appendix C for more details. Now, it is expected that the true regularization λ has the same expansion with γ converging to a different variable. The result in (28) leads us to

Result 4. The unknown expectations in Result 3 can be computed by

$$\mathcal{E}(\lambda^2) = \sigma^2 m(\ln m + (T-1)\ln\ln m + \mathcal{E}(\gamma) - \ln(T-1)!), \tag{29}$$

and

$$\mathcal{E}(\lambda) = \sigma\sqrt{m\ln m}\left\{1 + \frac{(T-1)\ln\ln m + \mathcal{E}(\gamma) - \ln(T-1)!}{2\ln m}\right\}, \tag{30}$$

where γ is a proper random variable and its expectation is a proper positive number.

The observations in Section V will show the accuracy of this approximation, using both a theoretical and empirical value of $\mathcal{E}(\gamma)$.

V. Numerical Results

In this section, we present some results supporting and illustrating the analysis in Sections IV-D and IV-C for the case of 2 sources.

A. Consistency

In Section IV-C, we analyzed the consistency of CLASS and identified the so called pure cases for which CLASS behaves consistently under the assumption of infinite estimation dynamic range. Equation (22) gives a condition for the DOA set to be such a case. We also showed that, asymptotically there exists a threshold function $\Delta\theta = \frac{\delta}{m}$ that defines such a consistency resolution. The exact value of δ can be found by (23). A simple numerical search shows that $\delta = 2.26\pi$ is the threshold. We remind that in practice, with a finite dynamic range and a discretized space, this upper bound can be reduced.

B. Asymptotic Performance

In Section IV-D, we computed the performance of CLASS in pure cases in many steps. Equations (25) and (24) connect the error covariance and the statistical bias to the two first moments of the optimal regularization parameter λ given in

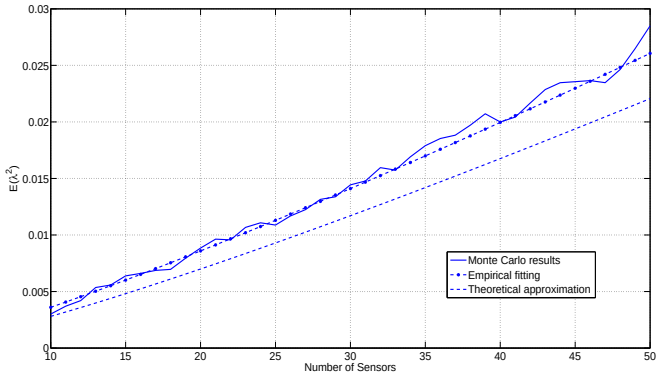

Fig. 1. The experimental value of $\mathcal{E}(\lambda^2)$ for different number of sensors compared to the theoretical results. The simple dashed line shows evaluation of (30) with $\mathcal{E}(\gamma) = 0.58$ while the dashed one with triangle shows the same evaluation with $\mathcal{E}(\gamma) = 1.3$.

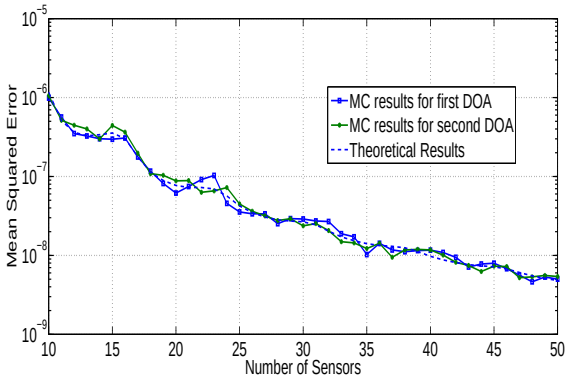

Fig. 2. The DOA MSE versus different number of sensors. The estimation is based on one snapshot measurement of two sources separated by $\Delta\theta = \frac{4\pi}{m}$, and waveform values $s_1 = s_2 = 1$, with the noise standard deviation $\sigma = 0.001$.

(26). The λ is then approximated to (27) for the asymptotic case of large m. Finally, (27) was further simplified to (28), assuming uncorrelated noise and almost constant number of snapshots. Here, we compare these results to the experimental ones. To find out the best estimation, we need the best λ value. Note that while theoretically sufficient, (26) is not applicable, since it is actually a first nonzero term of the exact best such value expanded for the high SNR case. Accordingly, we introduce the following procedure which works for high enough SNR.

Implementation

The optimization method is based on the homotopy rule and following the candidate points [14] starting from a large value of λ and no regressors ($n = 0$). For each $n > 0$, the corresponding candidate point can be found by the following iterative method recursively from the candidate point of $n-1$. First, fix λ and solve $\mathscr{P}_t$ only over the desired dimension of DOA n. This can be done precisely by a Newton method starting from the candidate point of $n-1$. Second, the solution found in Step 1 is the best solution if it is also the optimal solution for general $\mathscr{P}_t$ over all dimensions, but not with an infinitesimal decreasing of λ. This means that the second condition in (14) should hold with an extra global minimum. Thus, find the greatest peak p of $\|\mathbf{a}^H(\theta)\hat{\mathbf{N}}\|$ other than the estimated DOAs and update λ to $\frac{\lambda+p}{2}$. Iterate the first and second steps until convergence to where the next candidate point happens. We may continue this candidate point selection until the correct number of n has been reached.

As explained in Section IV-D, the first guess for $\mathcal{E}(\gamma)$ in (30) and (29) is the expected value of a normalized Gumble random variable, the Euler-Mascheroni constant [30]. In Figure 1, the simple dashed line with shows the theoretical value in (29) assuming the Gumble distribution. A better approximation can be found by fitting the experimental estimates, which gives $\mathcal{E}(\gamma) = 1.3$. Note that this is a fundamental constant for every DOA estimate by a half-wavelength ULA. In Figure 1, then, the dashed line with triangles shows such a result.

Fixing $\mathcal{E}(\gamma) = 1.3$, and combining (30) and (29) with (24) and (25) respectively, we complete evaluating the theoretical

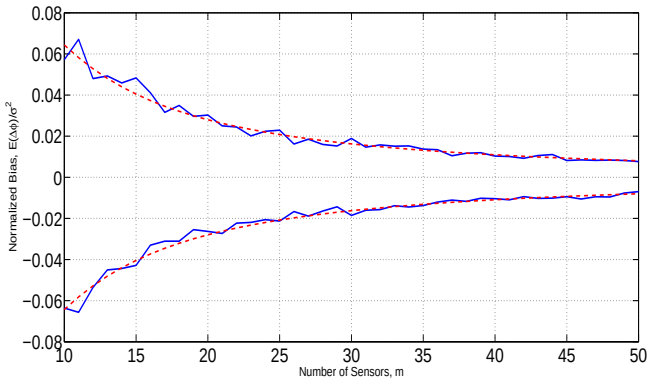

Fig. 3. The statistical bias normalized by true noise standard deviation for different number of sensors. The dashed line is the theoretical result. The estimation is based on one snapshot measurement of two sources separated by $\Delta\theta = \frac{4\pi}{m}$, and waveform values $s_1 = s_2 = 1$.

bias and covariance. Figures 2 and 3 show the performance of the estimator for different number of sensors. Note that the source separation $\Delta\theta$ changes by the number of sources as $\Delta\theta = \frac{4\pi}{m}$. We obtain the experimental results by using fixed values of the sources $s_1 = s_2 = 1$ and 1000 different realizations of the noise vector. The noise variance is fixed to $\sigma^2 = 10^{-6}$. However, since the performance is proportional to the noise level, the normalized results are shown. The results show agreement between the theory and the experimental results.

We then compared the LASSO performance to the ML and conventional beamforming ones. Figures 4 and 5 compare the estimation Mean Squared Errors and variances of three different estimators; CLASS, ML and conventional beamforming, respectively. The results show that while the asymptotic variances of CLASS and ML methods coincide, the CLASS estimator has a higher asymptotic MSE. We conclude that CLASS modifies the solution of ML mostly by adding a bias term in the very high SNR regime. However, as SNR decreases, the MSE of CLASS reaches the one for the ML estimator in the SNR regime between -2 and 5 dBs. The two methods reach the threshold region at almost the same SNR.

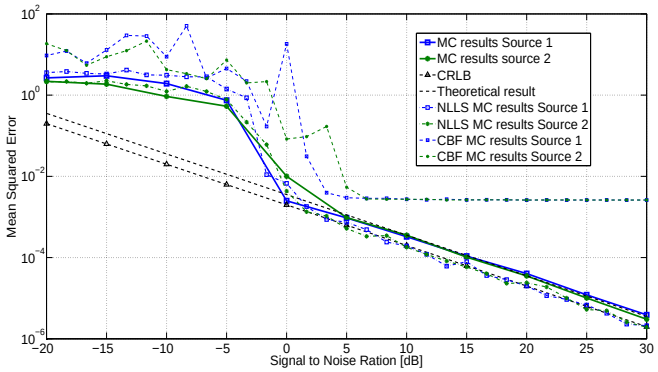

Fig. 4. The statistical MSE for different methods versus input SNRs. The estimation is based on one snapshot measurement of two sources separated by $\Delta\theta = \frac{4\pi}{m}$, and waveform values $s_1 = s_2 = 1$.

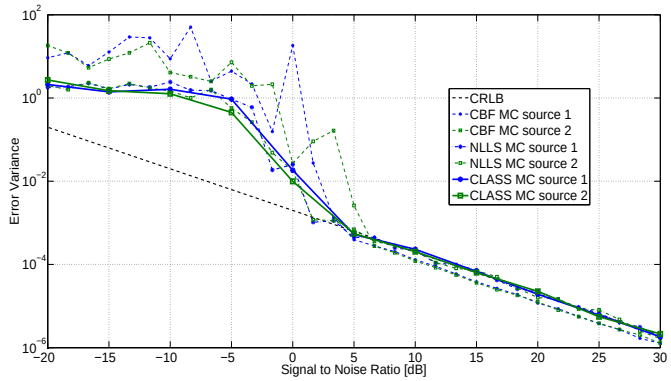

Fig. 5. The statistical variance for different methods in different input SNRs. The estimation is based on one snapshot measurement of two sources separated by $\Delta\theta = \frac{4\pi}{m}$, and waveform values $s_1 = s_2 = 1$.

VI. Conclusion

In this work, we studied the behavior of LASSO-based signal parameter estimation in terms of bias and variance. We remind that the aim of this work is to analyze a previously introduced method and not to devise a new technique. However, the analysis needs extra elaboration due to the improper definition of LASSO for statistical analysis. Accordingly, the concept of LASSO-based estimation was first clarified by introducing CLASS, a generalization of LASSO, which acts over a continuous set of indexes Θ rather than the finite matrix indexes. Then, we introduced conditions and definitions guaranteeing a good performance. Finally, we gave the theoretical performance under such consistency conditions. We finally tested our results by introducing an implementation of CLASS in a high SNR scenario. The results were confined to the worst case of the one snapshot scenario.

From the theoretical and experimental results, one may conclude that the LASSO technique sacrifices performance only in terms of statistical bias with the same threshold SNR to attain a better implementation as we explained in Section V. Note that our implementation of ML by searching for the best initial point is of exponential complexity if the number of sources increase, while the computational cost of our LASSO implementation grows linearly with n^3 due to the Newton method at each stage. This is of great interest due to the flexibility of LASSO with the choice of the manifold, which may vary due to the specific problem. Note also that the complex-valued nature of the problem limits the implementation techniques, in which case the current technique should be considered as an effective general solver.

Last but not least, we address the RP selection issue. As shown in Section C, the performance of CLASS is improved by choosing a smaller value of λ. Accordingly, one may propose implementing the noiseless LASSO, which theoretically guarantees the performance of the exact ML implementation. However, by introducing noise, some outliers will be introduced to the latter solution, which may be removed by thresholding. While unproven, we expect this thresholding strategy to have a higher threshold SNR, which brings the trade-off between the threshold and bias to account.

Appendix A
Uniqueness

Here, we give a proof for uniqueness of the LASSO based estimator. We start by bounding the dimension of the global optimum as in Theorem 3 which is an interesting and useful result by itself. Next, we prove the uniqueness based on this bound.

Lemma 1. For each matrix $\hat{\mathbf{N}}$, assume $\|\mathbf{a}^H(\theta)\hat{\mathbf{N}}\|_2 \leq \lambda$, where the steering vectors $\mathbf{a}(\theta)$ are of dimension m. Then, there exists at most $m-1$ points θ_i for which $\|\mathbf{a}^H(\theta_i)\hat{\mathbf{N}}\|_2 = \lambda$.

Proof: Take $\theta = \pi\cos(\theta)$. Then, for a ULA we have $\mathbf{a}(\theta) = [1 \;\; e^{j\theta} \;\; e^{j\theta} \;\; e^{j2\theta} \ldots e^{j(m-1)\theta}]^T$. We may also generalize this vector to the whole complex space as $\mathbf{a}(z) = [1 \; z \; z^2 \ldots z^{m-1}]^T$ where $z \in \mathbb{C}$. Obviously, $\mathbf{a}(\theta) = \mathbf{a}(e^{j\theta})$.

Define $F(\theta) = \|\mathbf{a}^H(\theta)\hat{\mathbf{N}}\|_2^2$, $\mathbf{T} = \hat{\mathbf{N}}\hat{\mathbf{N}}^H$, and $T_\alpha = \sum_{j-i=\alpha-m+1} T_{ij}$ for $0 \leq \alpha \leq 2(m-1)$. Furthermore, define the polynomial $S(z) = \sum_{\alpha=0}^{2(m-1)} T_\alpha z^\alpha - \lambda z^{m-1}$. It is straightforward to show that $F(\theta) = \frac{S(z)}{z^{m-1}} + \lambda$ for $z = e^{-j\theta}$. Note that since $F(z) = \frac{S(z)}{z^{m-1}} + \lambda$ is a complex differentiable function, from the Cauchy-Riemann equations we get $\frac{dF}{dz} = -\frac{1}{jz}\frac{\partial F}{\partial \theta}$ if $z = e^{-j\theta}$. Suppose now that $\boldsymbol{\theta} = [\theta_1 \;\; \theta_2 \ldots \theta_n]$ is a global minimum of $\mathscr{P}_t$. Then, due to the assumption, the function $F(\theta)$ is always less than or equal to λ and it has local maxima at $\theta = \theta_i$, with $F(\theta_i) = \lambda$. Then, $F(z_i) = 0$ and $\frac{dF}{dz}(z_i) = -\frac{1}{jz_i}\frac{\partial F}{\partial \theta}(\theta) = 0$, where $z_i = e^{-j\theta_i}$, with electrical angles θ_i corresponding to θ_i. It can simply be concluded that $S(z_i) = S'(z_i) = 0$, which means that all numbers z_i are multiple roots of $S(z)$. However, $S(z)$ is of degree $2(m-1)$ and the number of multiple zeros can not be more than $\frac{2(m-1)}{2} = m-1$. ∎

We also use the result in Theorem 3 in [25], Paper III. for the noiseless part. To prove the uniqueness, assume there exists two irreducible global minima $(\boldsymbol{\theta}_1, \mathbf{S}_1)$ and $(\boldsymbol{\theta}_2, \mathbf{S}_2)$. Define $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \cup \boldsymbol{\theta}_2$ and extend $\mathbf{S}_1$ and $\mathbf{S}_2$ on $\boldsymbol{\theta}$ to $\mathbf{S}^1$ and $\mathbf{S}^2$ as

$$\mathbf{S}_k^i = \begin{cases} \mathbf{S}_{i,l} & \theta_k = \theta_{i,l} \\ \mathbf{0} & \theta_k \notin \boldsymbol{\theta}_i \end{cases} \tag{31}$$

Note that the pairs $(\boldsymbol{\theta}, \mathbf{S}^i)$ are reducible global minima. Since $\Psi(\mathbf{X}, \boldsymbol{\theta}, \mathbf{S})$ is a convex function of $\mathbf{S}$, fixing λ, we have

$$\Psi(\mathbf{X}, \boldsymbol{\theta}, (1-\mu)\mathbf{S}^1 + \mu\mathbf{S}^2) \leq (1-\mu)\Psi(\mathbf{X}, \boldsymbol{\theta}, \mathbf{S}^1) + \mu\Psi(\mathbf{X}, \boldsymbol{\theta}, \mathbf{S}^2) = \Psi(\mathbf{X}, \boldsymbol{\theta}, \mathbf{S}^1) \quad (32)$$

for every $0 \leq \mu \leq 1$. For such μ, this means that the point $\left(\boldsymbol{\theta}, (1-\mu)\mathbf{S}^1 + \mu\mathbf{S}^2\right)$ is also a global minimum. Note also that the two terms $\frac{1}{2}\|\cdot\|_F^2$ and $\|\cdot\|_{1,2}$ are convex, thus

$$\|(1-\mu)\mathbf{S}^1 + \mu\mathbf{S}^2\|_{1,2} = (1-\mu)\|\mathbf{S}^1\|_{1,2} + \mu\|\mathbf{S}^2\|_{1,2} \quad (33)$$

and

$$\tfrac{1}{2}\left\|\mathbf{X} - \mathbf{A}(\boldsymbol{\theta})\left((1-\mu)\mathbf{S}^1 + \mu\mathbf{S}^2\right)\right\|_F^2 = (1-\mu)\tfrac{1}{2}\left\|\mathbf{X} - \mathbf{A}(\boldsymbol{\theta})\mathbf{S}^1\right\|_F^2 + \mu\tfrac{1}{2}\left\|\mathbf{X} - \mathbf{A}(\boldsymbol{\theta})\mathbf{S}^2\right\|_F^2 \quad (34)$$

since, otherwise, (34) and (33) with strict inequality results in strict inequality in (32). Moreover, $\frac{1}{2}\|\cdot\|_F^2$ is a strictly convex function and from (34) we conclude that $\mathbf{A}(\boldsymbol{\theta})(\mathbf{S}_1 - \mathbf{S}_2) = \mathbf{0}$. But, this can be possible only if the dimension of $\boldsymbol{\theta}$ is more than m, since for $n = m$ the matrix $\mathbf{A}(\theta)$ is a Vandermonde type matrix and is full-rank. Choose μ such that $\frac{\mu}{1-\mu} < \frac{\|\mathbf{S}_i^1\|_2}{\|\mathbf{S}_i^2\|_2}$ for every i. Then the pair $(\boldsymbol{\theta}, (1-\mu)\mathbf{S}^1 + \mu\mathbf{S}^2)$ is irreducible with $n > m$. This is a contradiction, since due to Corollary 1 of Theorem 2 and Lemma 1, the dimension n of every irreducible global minimum $\boldsymbol{\theta}$ of $\mathscr{P}_t$, is less than the dimension of the received vector m.

Finally, we prove the uniqueness of $\mathscr{P}_{t,\text{nl}}$. Note that the number of active DOAs for each global minimum of $\mathscr{P}_{t,\text{nl}}$ is also less than $m-1$. To show that, we assume a solution $(\boldsymbol{\theta}, \mathbf{S})$ with $n > m-1$. Take an arbitrary DOA θ and note that since $n > m-1$ the matrix $\mathbf{A} = [\mathbf{A}(\boldsymbol{\theta})\ \mathbf{a}(\theta)]$ has a nonempty null space. Then, we can take an $(n+1)\times T$ matrix $\mathbf{M}_0 = [-\mathbf{M}\ \boldsymbol{\nu}]^T$ such that $\mathbf{A}\mathbf{M}_0 = \mathbf{0}$. Note that every element of each vector in $\mathscr{N}_{\mathbf{A}}$ is nonzero, since every submatrix of $\mathbf{A}$ by removing columns is full rank. Thus, $\boldsymbol{\nu}$ can be made an arbitrary vector by scaling. From Theorem 3 in [25], Paper III., We have

$$-\mathbf{A}(\boldsymbol{\theta})\mathbf{M} + \mathbf{a}(\theta)\boldsymbol{\nu} = 0 \rightarrow \text{Re}(\text{Tr}(\boldsymbol{\Gamma}_{\mathbf{S}}^{-1}\mathbf{S}\mathbf{M}^H)) + \|\boldsymbol{\nu}\|_{1,2} \geq 0 \quad (35)$$

Note also that, using Corollary 4 in [25], Paper III., there exists a vector $\mathbf{Z}$ such that $\mathbf{A}(\theta)^H\mathbf{Z} = \boldsymbol{\Gamma}_{\mathbf{S}}^{-1}\mathbf{S}$. Combining this with (35), we get $\text{Re}(\text{Tr}(\mathbf{M}^H\mathbf{A}^H\mathbf{Z})) + \|\boldsymbol{\nu}\|_2 \geq 0$. Recalling that $\mathbf{A}\mathbf{M} = \mathbf{a}(\theta)\boldsymbol{\nu}$ we have $\text{Re}(\text{Tr}(\boldsymbol{\nu}^H\mathbf{a}^H(\theta)\mathbf{Z})) + \|\boldsymbol{\nu}\|_2 \geq 0$ for every $\boldsymbol{\nu}$. It is easy to verify that this is true if and only if $\|\mathbf{a}^H(\theta)\mathbf{Z}\|_2 \leq 1$. Note that from the definition of $\mathbf{Z}$, we have $\|\mathbf{a}^H(\theta_i)\mathbf{Z}\|_2 = 1$. As the conditions of Lemma 1 are satisfied, we conclude that $n < m$. Like the previous part, the assumption of two global minima leads us to a third one as a convex combination of the two with $n > m$ which is a contradiction.

APPENDIX B
CONTINUOUS SOLUTION PATH

First, note that for each pair $(\mathbf{X}, \lambda)$ and $\epsilon > 0$, there exists an $\alpha > 0$ such that if $\Psi(\mathbf{X}, \boldsymbol{\theta}, \mathbf{S}, \lambda) \leq P(\mathbf{X}, \hat{\boldsymbol{\theta}}(\mathbf{X}, \lambda), \lambda) + \alpha$ and $\boldsymbol{\theta}$ is of bounded dimension, then for all $\theta \in \boldsymbol{\theta}$ either there exists a $\theta' \in \hat{\boldsymbol{\theta}}(\mathbf{X}, \lambda)$ so that $\max\{|\theta - \theta'|, \|\mathbf{S}_\theta - \mathbf{S}_{\theta'}\|\} < \epsilon$, or $\|\mathbf{S}_{\theta'}\| < \epsilon$. Otherwise, using the BolzanoWeierstrass theorem, there exists a sequence $\boldsymbol{\theta}_k, \mathbf{S}_k$ converging to a global $\mathscr{P}_t$ point different from $\hat{\boldsymbol{\theta}}, \hat{\mathbf{S}}$ which is a contradiction.

Next, note that due to Theorem 3 every $\mathscr{P}_t$ solution dimension is bounded by m, and $\mathbf{S}$ can simply be confined to a sufficiently big but compact set e.g. $\|\mathbf{S}\|_{1,2} \leq A$. Over this set, as the pair $(\mathbf{X}', \lambda')$ converge to $(\mathbf{X}, \lambda)$, the continuous function $\Psi_{\mathbf{X}',\lambda'}(\boldsymbol{\theta}, \mathbf{S}) = \Psi(\mathbf{X}', \boldsymbol{\theta}, \mathbf{S}, \lambda')$ uniformly converges to $\Psi_{\mathbf{X},\lambda}(\boldsymbol{\theta}, \mathbf{S})$. This means that there exists a $\delta > 0$ such that if $\max\{\|\Delta\mathbf{X}\|, \|\Delta\lambda\|\} < \delta$, then for every proper pair $(\boldsymbol{\theta}, \mathbf{S})$ we have $|\Psi_{\mathbf{X}+\Delta\mathbf{X},\lambda+\Delta\lambda}(\boldsymbol{\theta}, \mathbf{S}) - \Psi_{\mathbf{X},\lambda}(\boldsymbol{\theta}, \mathbf{S})| < \frac{\alpha}{2}$. This leads to $\Psi_{\mathbf{X},\lambda}(\hat{\boldsymbol{\theta}}(\mathbf{X}+\Delta\mathbf{X}, \lambda+\Delta\lambda), \hat{\mathbf{S}}(\mathbf{X}+\Delta\mathbf{X}, \lambda+\Delta\lambda)) \leq \Psi_{\mathbf{X},\lambda}(\hat{\boldsymbol{\theta}}(\mathbf{X}, \lambda), \hat{\mathbf{S}}(\mathbf{X}, \lambda)) + \alpha$. Using the results above the proof is complete.

To extend the homotopy to $\lambda = 0$, note that from the minimality of $\hat{\boldsymbol{\theta}}(\mathbf{X}, \lambda_k)$, we conclude

$$\frac{1}{2}\left\|\mathbf{X} - \mathbf{A}(\hat{\boldsymbol{\theta}}(\mathbf{X}, \lambda))\hat{\mathbf{S}}(\lambda)\right\|_F^2 + \lambda\|\hat{\mathbf{S}}(\lambda)\|_{1,2} \leq \lambda\|\hat{\mathbf{S}}_{\text{nl}}\|_{1,2} \quad (36)$$

for all $\lambda > 0$. Now, the continuity holds. Otherwise, there exists a sequence $\{\lambda_k\}_{k=1}^{\infty}$ of regularization parameters converging to zero, for each of which there exists $\theta \in \hat{\boldsymbol{\theta}}(\lambda)$ so that $\max\{|\theta - \theta'|, \|\mathbf{S}_\theta - \mathbf{S}_{\theta'}\|\} > \epsilon$ and $\|\mathbf{S}_\theta\| > \epsilon$ for every $\theta' \in \hat{\boldsymbol{\theta}}_{\text{nl}}$. From (36) and the uniqueness theorem, we conclude that the solutions are in a compact set of bounded dimension. Thus, there exists a subsequence of λ_k for which $\hat{\boldsymbol{\theta}}(\mathbf{X}, \lambda_k)$ and $\hat{\mathbf{S}}(\lambda_k)$ converge to some $\boldsymbol{\theta}$ and $\hat{\mathbf{S}}$ respectively, different from the noiseless solutions. From (36), we get $\frac{1}{2}\left\|\mathbf{X} - \mathbf{A}(\hat{\boldsymbol{\theta}}(\mathbf{X}, \lambda))\hat{\mathbf{S}}(\lambda)\right\|_F^2 \leq \lambda\|\hat{\mathbf{S}}_{\text{nl}}\|_{1,2}$, and letting k tend to infinity, we conclude $\mathbf{A}(\boldsymbol{\theta})\hat{\mathbf{S}} = \mathbf{X}$. On the other hand, $\lambda_k\|\hat{\mathbf{S}}(\lambda_k)\|_{1,2} \leq \lambda_k\|\hat{\mathbf{S}}_{\text{nl}}\|_{1,2}$, and we get $\|\hat{\mathbf{S}}\|_{1,2} \leq \|\hat{\mathbf{S}}_{\text{nl}}\|_{1,2}$. This is in contradiction with the uniqueness of the noiseless solution.

APPENDIX C

A. Computing the error covariance

We start by $\Delta\boldsymbol{\theta} = \lambda\boldsymbol{\beta} + \boldsymbol{\delta}$. Note that since $\boldsymbol{\delta}$ is the first order ML error, its covariance can be found, e.g. in [27]. We get $\text{Cov}(\Delta\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\Gamma}^{-1}\mathbf{R}^{-1}\text{Re}[\mathbf{D}_p^H\mathbf{C}\mathbf{D}_p \odot \boldsymbol{\Xi}^T]\mathbf{R}^{-1}\boldsymbol{\Gamma}^{-1} + \text{Var}(\lambda)\boldsymbol{\beta}\boldsymbol{\beta}^T + \mathbf{k}$ where $\mathbf{k}$ is proportional to $\mathcal{E}(\lambda\Delta\mathbf{X})$. Now, we show that this expectation is zero and the effective terms are related to the second order statistics of λ.

We can write the best λ value in (26) as

$$\lambda = \min\left\{a | a \geq \max_{\theta}\|\mathbf{a}^H(\theta)[\boldsymbol{\Upsilon} + a\boldsymbol{\Lambda}]\|_2\right\} \quad (37)$$

where $\boldsymbol{\Upsilon}$ is a linear (but not analytic) matrix function of $\Delta\mathbf{X}$ and $\boldsymbol{\Lambda}$ is a constant one. Thus, $\boldsymbol{\Upsilon}$ is centered Gaussian but not circularly symmetric. We may show λ as $\lambda(\boldsymbol{\Upsilon})$ or $\lambda(\Delta\mathbf{X}) = \lambda(\boldsymbol{\Upsilon}(\Delta\mathbf{X}))$.

The first observation is that $\lambda(\alpha\boldsymbol{\Upsilon}) = |\alpha|\lambda(\boldsymbol{\Upsilon})$ for every complex α, which implies that $\lambda(\alpha\Delta\mathbf{X}) = |\alpha|\lambda(\Delta\mathbf{X})$. In other words, for every $a > 0$, $\{\Delta\mathbf{X}|\lambda(\Delta\mathbf{X}) = a\} = \{a\Delta\mathbf{X}|\lambda(\Delta\mathbf{X}) = 1\}$ which means that $\mathcal{E}(\Delta\mathbf{X}|\lambda = a) = a\mathcal{E}(\Delta\mathbf{X}|\lambda = 1)$. We conclude that $\mathcal{E}(\lambda)\mathcal{E}(\Delta\mathbf{X}|\lambda = 1) = \mathcal{E}_\lambda(\mathcal{E}(\Delta\mathbf{X}|\lambda)) = \mathcal{E}(\Delta\mathbf{X}) = \mathbf{0}$. Thus, $\mathcal{E}(\Delta\mathbf{X}|\lambda = 1) = 0$ and

$\mathcal{E}(\Delta\mathbf{X}\lambda) = \mathcal{E}_\lambda(\mathcal{E}(\lambda\Delta\mathbf{X}|\lambda)) = \mathcal{E}(\lambda^2)\mathcal{E}(\Delta\mathbf{X}|\lambda = 1) = 0$. This shows that $\mathbf{k} \propto \mathcal{E}(\lambda\Delta\mathbf{X}) = 0$.

B. The asymptotic extreme value expansion

The variables z_k introduced in Section IV-D are independent with chi squared distribution. Thus,

$$\Pr(z_i > \alpha) = \frac{1}{(T-1)!}\int_{\frac{\alpha}{\sigma^2 m}}^{\infty} z^{T-1}e^{-z}\mathrm{d}z \qquad (38)$$

which is in the order of $\frac{\alpha^{T-1}e^{-\frac{\alpha}{\sigma^2 m}}}{(T-1)!(\sigma^2 m)^{T-1}}$. This can be seen , for example, by L'Hopital's rule. Since z_is are independent, it is simple to see that

$$\Pr(\lambda_s^2 < \alpha) = (1 - \Pr(z_i > \alpha))^m = \beta \qquad (39)$$

Then, for large enough m, we get $\Pr(z_i > \alpha) = \frac{-\ln\beta}{m}$ which using (38), can be written as $(T-1)\ln\frac{\alpha}{\sigma^2 m} - \frac{\alpha}{\sigma^2 m} - \ln(T-1)! = \ln(-\ln(\beta)) - m$. It can then be simplified to

$$\frac{\alpha}{\sigma^2 m} = \ln m + (T-1)\ln\ln m + \gamma - \ln(T-1)! + o(1) \quad (40)$$

where $\gamma = -\ln(-\ln\beta)$. We conclude (28).

REFERENCES

[1] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, vol. 25, no. 6, pp. 714–721, Mar. 2009.

[2] W. Tu and S. Sun, "Spatial filter selection with lasso for eeg classification," in *Advanced Data Mining and Applications*, ser. Lecture Notes in Computer Science, L. Cao, J. Zhong, and Y. Feng, Eds. Springer Berlin / Heidelberg, 2010, vol. 6441, pp. 142–149.

[3] Z. Wang and S. Tan, "Identifying idiosyncratic stock return indicators from large financial factor set via least angle regression," *Expert Systems with Applications*, vol. 36, no. 4, pp. 8350 – 8355, 2009.

[4] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B, (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[5] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, Dec. 1998.

[6] J.-J. Fuchs, "On the application of the global matched filter to doa estimation with uniform circular arrays," *Signal Processing, IEEE Transactions on*, vol. 49, no. 4, pp. 702 –709, april 2001.

[7] M. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 1, no. 4, pp. 586 –597, 2007.

[8] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[9] T. Park and G. Casella, "The bayesian lasso," *Journal of the American Statistical Association*, vol. 103, p. 681 686, 2008.

[10] H. Zayyani, M. Babaie-Zadeh, and C. Jutten, "Bayesian pursuit algorithm for sparse representation," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, april 2009, pp. 1549 –1552.

[11] D. Donoho and Y. Tsaig, "Fast solution of -norm minimization problems when the solution may be sparse," *Information Theory, IEEE Transactions on*, vol. 54, no. 11, pp. 4789 –4812, nov. 2008.

[12] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004.

[13] S. Wright, R. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," *Signal Processing, IEEE Transactions on*, vol. 57, no. 7, pp. 2479 –2493, july 2009.

[14] A. Panahi and M. Viberg, "Fast candidate points selection in the lasso path," *Signal Processing Letters, IEEE*, vol. 19, no. 2, pp. 79 –82, feb. 2012.

[15] E. J. Candes and Y. Plan, "Near-ideal model selection by $\ell_1$minimization," *Annals of Statistics*, vol. 37, p. 2145, 2009.

[16] D. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289 –1306, april 2006.

[17] E. J. Candes and Y. Plan, "A Probabilistic and RIPless Theory of Compressed Sensing," *Information Theory, IEEE Transactions on*, vol. 57, no. 11, pp. 7235–7254, Nov. 2011.

[18] D. Malioutov, M. Cetin, and A. Willsky, "Source localization by enforcing sparsity through a laplacian prior: an svd-based approach," *IEEE Workshop on Statistical Signal Processing*, pp. 573 – 576, sept.-1 oct. 2003.

[19] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," http://cvxr.com/cvx, Apr. 2011.

[20] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *Signal Processing Magazine, IEEE*, vol. 13, no. 4, pp. 67 –94, jul 1996.

[21] A. Panahi and M. Viberg, "Maximum aposteriory based regularization parameter selection," in *International Conference on Acoustics, Speech, and Signal Processing*, 2011.

[22] E. J. Candes and T. Tao, "Decoding by linear programming," *Information Theory, IEEE Transactions on*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[23] P. Stoica and P. Babu, "Sparse estimation of spectral lines: Grid selection problems and their solutions," *Signal Processing, IEEE Transactions on*, vol. 60, no. 2, pp. 962 –967, feb. 2012.

[24] M. R. Osborne, B. Presnell, and B. Turlach, "A new approach to variable selection in least squares problems," 1999.

[25] A. Panahi, "Parameter estimation using sparse modeling: Algorithms and performance analysis," *Licentiate thesis, Chalmers University of Technology, Department of Signals and Systems*, September 2012.

[26] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *Information Theory, IEEE Transactions on*, vol. 47, no. 7, pp. 2845 –2862, nov 2001.

[27] M. Viberg, B. Ottersten, P. Stoica, and A. Nehorai, "Exact and large sample ml techniques for parameter estimation and detection in array processing," in *Radar Array Processing*, T. J. S. Simon S. Haykin, John Litva, Ed. Springer-Verlag, 1993, ch. 4.

[28] A. Panahi and M. Viberg, "On the resolution of the lasso-based doa estimation method," in *Smart Antennas (WSA), 2011 International ITG Workshop on*, feb. 2011, pp. 1 –5.

[29] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels, *Statistics of Extremes: Theory and Applications*, 1st ed. Wiley, Oct. 2004.

[30] D. Knuth, "Eulers constant to 1271 places," *Math. Comput.*, vol. 16, no. 275-281, 1962.